Introduction to Survey Weights for 2009-2010 National Adult Tobacco Survey

Sean Hu, MD., MS., DrPH
Office on Smoking and Health
Presented to Webinar
January 18, 2012

Outline

• Overview of design and weighting methodology

• Weighting procedures

• Final weights
OVERVIEW OF DESIGN AND WEIGHTING METHODOLOGY

Major survey errors with a random-digit-dialing survey

Coverage

Sampling

Nonresponse

Measurement (self-reported data)
Percentage of U.S. Households Without Landline Telephones

- Based on National Health Interview Survey data
A Dual-Frame RDD Survey

• RDD expands a traditional landline based RDD survey to a dual frame survey of landline and cell phone numbers

• Reach 98% of US households
Landline and Cell phone populations and frames

Non-overlapping dual frames:
- Landline RDD frame = A + B
- Cell phone-only frame = D

RDD Sampling
Disproportionate Stratified Sampling

DSS – Landline telephone numbers are classified into strata that are either high density (listed 1+ block telephone numbers) or medium density (not listed 1+ block telephone numbers) to yield residential telephone numbers. The sampling ratio is 1.5:1

NATS sampling frame
- Landline sampling frame: a list landline stratum
  a not-listed landline stratum
- Cell phone stratum

*Citations, references, and credits – Myriad Pro, 11pt*
### NATS final disposition code distribution

<table>
<thead>
<tr>
<th>Categories of final disposition codes</th>
<th>Landline</th>
<th></th>
<th>Cell Phone</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>Complete &amp; partial complete</td>
<td>110,634</td>
<td>5.5</td>
<td>7,947</td>
<td>2.0</td>
</tr>
<tr>
<td>Eligible but not interview</td>
<td>93,217</td>
<td>4.6</td>
<td>4,112</td>
<td>1.0</td>
</tr>
<tr>
<td>Unknown eligibility, non-interview</td>
<td>519,773</td>
<td>25.6</td>
<td>247,905</td>
<td>62.5</td>
</tr>
<tr>
<td>Not eligible</td>
<td>1,303,822</td>
<td>64.3</td>
<td>136,932</td>
<td>34.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,027,446</strong></td>
<td><strong>100.0</strong></td>
<td><strong>396,896</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>

There were many ineligible frame members and members with unknown eligible status.

---

### Demographic Distributions between 2010 Census and 2009-2010 NATS

<table>
<thead>
<tr>
<th>Gender</th>
<th>2010 Census</th>
<th>2009-2010 NATS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>Male</td>
<td>151781326</td>
<td>49.16</td>
</tr>
<tr>
<td>Female</td>
<td>156964212</td>
<td>50.84</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.15</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>2010 Census</th>
<th>2009-2010 NATS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>18-24</td>
<td>30672088</td>
<td>13.08</td>
</tr>
<tr>
<td>25-44</td>
<td>82134554</td>
<td>35.02</td>
</tr>
<tr>
<td>45-64</td>
<td>81489445</td>
<td>34.74</td>
</tr>
<tr>
<td>&gt;=65</td>
<td>40267984</td>
<td>17.17</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.15</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>2010 Census</th>
<th>2009-2010 NATS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>White Only, Non-Hispanic</td>
<td>196817552</td>
<td>63.75</td>
</tr>
<tr>
<td>Black Only, Non-Hispanic</td>
<td>37685848</td>
<td>12.21</td>
</tr>
<tr>
<td>Asian Only, Non-Hispanic</td>
<td>14465124</td>
<td>4.69</td>
</tr>
<tr>
<td>Other, Non-Hispanic</td>
<td>9299420</td>
<td>3.01</td>
</tr>
<tr>
<td>Hispanic</td>
<td>50477594</td>
<td>16.35</td>
</tr>
<tr>
<td>Unknown</td>
<td>1.4</td>
<td>1.22</td>
</tr>
</tbody>
</table>

Male, young adults, and minority were under-represented in the survey.
Some challenges after data collection:

- Having some coverage error:
  - Around 2% of US households don't have any phone.
- Declining response rate
- Low response rates with some specific group under-represented in the survey
  - Smoking rates are relatively higher among these groups.

What is weighting?

- Weighting is a process used to remove bias in the sample.
- Corrects for difference in the probability of selection due to non-response and non-coverage errors
- Adjust variables of age, race, gender, and other demographic and related variables between the sample and the entire population
- Allows the generalization of findings to the whole population.
Sample Weights

- Are assigned to each sample member
- Reflect differences between the distribution of the sample and the population
- Can be viewed as the number of population members that the sample unit represents

NATS WEIGHTING PROCEDURES

1. Calculate design weights
2. Adjust for the frame members with unknown eligibility status
3. Adjust for the people who did not respond to the survey (nonresponse adjustment).
4. Poststratification
Design Weights

- Stratum design: the inverse of the probability of selection (1/P)
- Multiply number of adults in household for landline sample (NUMADULT)
- The inverse of number of phones in household for landline sample (1/NUMPHONE)
- Design Weight = (1/P) * (1/NUMPHONE) * NUMADULT
  - P = Probability of selection
  - NUMPHONE = number of phones within the household
  - NUMADULT = number of adults eligible for the survey within the household

Adjustment for unknown eligibility status

- Purpose is to adjust sampled frame members with unknown eligibility status.
- Stratum weight = the inverse of the probability of selection at a stratum = Total frame members at a stratum/sample members
  - Sampled members includes completed, refused, non-contact (unknown eligibility status), and ineligible members.
  - Sampled members with unknown eligibility could be nonrespondents or they could be ineligible
- Adjusted factor = (completed + refused)/(completed + refused + ineligible)
  - It is the estimated percentage of known eligible members at a stratum
- Adjusted stratum weight = stratum weight * adjusted factor
Nonresponse Adjustment

• If every person sampled agreed to do the survey, then weighted estimates using just the design weights would be sufficient to estimate the population values.

• However, every survey has some level of nonresponse.

• We can adjust for this nonresponse by spreading the weight of the nonrespondents to the respondents.

Nonresponse Adjustment (continued)

• In order to perform this adjustment, we need the following:

  – The status of each interview:
    • Complete, partially complete
    • Refusal
    • Unable to contact
    • Ineligible

  – Characteristics used in the adjustment for all members of the sample including refusals, no contacts, and eligibles.
Step 1: Using Auxiliary Data

- A final adjustment can be made to bring the sample weights of who we did contact up to the level of the population.
- This is done when there are accurate population totals from a source other than the sample frame.
- This is a way to account for under- or overcoverage of certain demographic groups.
- For example, we frequently under-cover young adults aged 18-24. If we have good estimates of the population for this group, we can adjust the weighted total for the group up to the amount in the auxiliary data.

Using Auxiliary Data (Continued)

Auxiliary data (America Community Survey) at the various geographical levels was appended to the sample frame.

- Block group level information for listed landline sample
- County FIPS code information for unlisted sample
- Area code information for the cell sample

The following variables are used in the nonresponse adjustment procedure:
- Population density
- Proportion white
- Proportion African-American
- Proportion Hispanic
- Proportion of families below 150% of the poverty line
- Proportion that are high school graduates
- Proportion that completed a Bachelor’s degree
Step 2: Predict response propensity

- Auxiliary data needed to model sample units' response propensities
  - A logistic regression is used to obtain a probability of response ($\rho$) for every unit
    - the outcome is response and the independent variables are the auxiliary data

Step 3: A new weight is calculated

\[ W_{i,k,j}^2 = W_{i,k,j}^1 * \frac{1}{P_{i,k,j}} \]

- $W_{i,k,j}^2$ is the nonresponse adjusted probability for the $j$th unlisted frame member
- $P_{i,k,j}$ is the predicted probability of response for the $j$th unlisted respondent from the logistic model.

We also make the following ratio adjustment.

\[ W_{i,k,j}^3 = \frac{\sum W_{i,k,j}^1 W_{i,k,j}^2}{\sum W_{i,k,j}^2} \]
Poststratification

• Poststratification adjusts the sample to the target population to insure that the distribution of the sample aligns with the distribution in the population for some set of variables.

• To remove nonresponse and coverage bias

Selecting variables in Poststratification

There are three ways to select the variables:

• When using a dual frame we poststratify to the population totals for each phone type (cell-only, landline including dual user)

• Use some common demographic variables such as: age, gender, race/ethnicity, etc.

• Selecting variables that are most highly correlated with your outcome of interest such as education and marital status.
Poststratification

Several Possible Approaches:

1. Use a single big age x gender x education table for the calculation of the weights (traditional poststratification).
   - However, crosstabs may not be available for the population
   - and, small cell sizes in the sample table

2. Iterative Solutions:
   - Manual version (stepwise programming in statistical software)
   - Automatic version (i.e. Raking software)
   - Logistic regression based solutions
     - NATS used the logistic regression approach.

Imputation and Trimming

- The poststratification variables were imputed.

- Weight trimming is applied in order to constrain the most extreme weight, and thereby reduce variance.
Cell Phone Respondent Issue

- For state with very few cell phone respondents, including cell phone respondents would result in large unequal weighting effects, and consequently, large variance for population estimates.

- We chose 200 cell phone respondents as a cut-off point for including cell phone respondents at state level.

Correlation between unequal weighting effect (UWE) and number of cell phone respondents
Final Weights

• WT_NATIONAL:
  • for national estimates
  • use all of respondents (both landline and cell respondents) in all of the states

• WT_LANDLINE:
  • only for landline respondents
  • only use landline respondents

• WT_STATE:
  • for state estimates
  • for states with 200+ cell cases, use both respondents
  • For state with less 200 cell cases, only use landline cases

Question or comments?
Thanks!

Contact Info:
Sean Hu, MD, MS, DrPH
Telephone: 770-488-5845
E-mail: shu@cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.